Construction and Sharing Mechanism of Bioinformatics Databases: Data Governance Strategies for Cross-Species Comparative Genomics

Yu Ding
Jintan No.1 High School, Changzhou, China

Abstract: With the rapid development of high-throughput sequencing technology, the volume of bioinformatics data has shown explosive growth, and cross-species comparative genomics has become an important means to analyze the laws of life evolution, gene functions and disease mechanisms. However, data heterogeneity, privacy protection requirements and technical barriers have led to a serious phenomenon of data silos, restricting scientific research collaboration and innovation. This paper systematically explores the core issues of cross-species comparative genomics data governance from three dimensions: database architecture design, data governance strategies, and innovation of sharing mechanisms. It proposes solutions based on modular design, distributed storage, and dynamic governance frameworks, and verifies their feasibility by combining global typical cases, providing theoretical support for biomedical data integration and open science.

Keywords: Bioinformatics Database; Cross-Species Comparative Genomics; Data Governance; Distributed Storage; Privacy Protection.

1. Introduction

Genomics research has entered the era of "multi-omics big data", and the amount of data generated by technologies such as single-cell sequencing, three-dimensional genomics and metagenomics has grown exponentially. For example, in the Human Pangenome Project, HPP is expected to generate more than 1PB of genomic variation data [1], while the Earth Genome Project [2] aims to complete the genome sequencing of 1.5 million species [3]. However, currently approximately 70% of bioinformatics data are scattered and stored in different laboratories or institutions, resulting in problems such as repeated sequencing, resource waste, and non-reproducibility of research results [4].

Cross-species comparative genomics reveals the molecular basis of gene family evolution, epigenetic regulatory mechanisms and species adaptive evolution through the comparison of genomes of multiple species. For instance, by comparing the expression profiles of neurodevelopmental genes in humans and zebrafish, it was found that the core gene modules regulating neuronal migration are highly conserved in vertebrates, providing a new target for autism research [5]. However, such research highly relies on the standardized integration of cross-species data, while existing databases have significant deficiencies in aspects such as data format, metadata management, and access control [6].

2. Research Background and Problem Presentation

2.1 Explosive Growth and Isolation of Data

Illumina NovaSeq X Plus has reduced the cost of whole-genome sequencing to less than $200 [7] and driven the annual growth rate of data volume to exceed 40%. Approximately 85% of the global genomic data is stored on local servers, and only 15% is shared through public databases [8]. Repeated sequencing leads to a waste of approximately 120 million US dollars of funds each year [2].

2.2 Requirements and Challenges of Cross-Species Comparative Genomics

Through comparative genomics, it has been found that approximately 80% of disease-related genes in humans and mice are conserved [9].

Multi-species genomic alignment needs to address issues such as differences in reference genomes and inconsistent annotations of

repetitive sequences [10]. The existing databases lack cross-species metadata standards, resulting in 70% of comparative genomics studies requiring manual data integration.

2.3 Research Objectives and Significance

This paper aims to construct a bioinformatics database framework that supports cross-species comparative genomics and proposes the following innovations:

Achieve functional decoupling through microservice design and support dynamic expansion; Combining blockchain with differential privacy technology to balance data sharing and privacy protection; Drawing on the experiences of EBP and HPP, a data sharing path suitable for developing countries is proposed.

3. Core Principles and Technical Paths for the Construction of Bioinformatics Databases

The construction of bioinformatics databases is the infrastructure of modern life science research, and it is necessary to balance the data scale, technical complexity and application requirements.

3.1 Core Principle: Build the Cornerstone of Data Governance

The construction of a bioinformatics database needs to follow six core principles. Data quality priority is the foundation of scientific research. Genomic sequencing data need to undergo multi-dimensional quality control, including sequencing error rate detection, continuity assessment of genomic assembly, and verification of variant sites, to ensure the reliability of the analysis results. Data security and privacy protection are the top priorities for bioinformatics databases. It is necessary to combine encrypted storage, dynamic access control and cross-border transmission compliance review to build a security protection system covering the entire data life cycle. Scalability and modular design are the keys to coping with future technological iterations. Databases need to support distributed storage and microservice development, and achieve environment consistency and elastic scalability through containerized deployment. Standardization and interoperability are the prerequisites for breaking down data silos. It is necessary to unify the format of genomic data and metadata standards, and provide open API interfaces to support cross-platform data sharing. User experience optimization can significantly lower the technical threshold. Through graphical interfaces, intelligent search and automated analysis processes, researchers can complete complex data retrieval and mining without programming. Performance optimization focuses on the efficiency of large-scale data queries. Through indexing technology, caching mechanisms and asynchronous processing, the response time of time-consuming tasks such as genome alignment is shortened to an acceptable range. These principles support each other and jointly constitute the guiding framework for the construction of bioinformatics databases.

3.2 Technical Implementation Path: The Entire Chain from Data Collection to Analysis

The construction of a bioinformatics database needs to cover the entire process of data collection, storage, processing and sharing. The data collection stage needs to support the integration of multi-source data, including synchronizing data from public databases and uploading raw data from local sequsequers, and achieving quality assessment and decontamination treatment through an automated cleaning assembly line. Cross-species genomic alignment and analysis is one of the core functions. It is necessary to combine short-read and long-read sequencing technologies, achieve precise matching of sequencing data with reference genomes through efficient alignment tools, and further explore genomic collinearity and structural variations. The integration of three-dimensional genomic data requires solving the problems of cross-species coordinate alignment and TAD boundary alignment, and revealing conserved regulatory elements and species-specific structural variations through visualization tools. Cross-species analysis of single-cell data focuses on batch effect correction, pseudo-temporal trajectory construction and gene regulatory network inference to reveal the conserved mechanisms of cell types and developmental processes. In terms of data sharing and collaboration, it is necessary to combine the federated learning framework with blockchain evidence preservation technology to achieve multi-institutional joint modeling and trusted traceability under the premise of safeguarding data sovereignty. At the same time, a visual collaboration platform that supports real-time editing by multiple people

should be developed to enhance the collaborative efficiency of cross-regional teams. This technical path runs through the entire data life cycle and provides support for the practical application of bioinformatics databases.

3.3 Typical Application Scenarios: From Basic Research to Clinical Transformation

The construction of bioinformatics databases needs to be closely combined with actual demands and play a key role in multiple fields. In the research of human disease genomics, the construction of cross-species disease models and the discovery of drug targets are important directions. By comparing the disease-related gene variations in humans and model organisms, conserved pathogenic mechanisms can be revealed and potential drug targets can be screened. In agricultural species improvement, the mining of stress-resistant genes and the prediction of heterosis rely on cross-species genomic alignment data. By identifying the differential genes between wild species and cultivated species, it can guide the improvement of crop stress resistance and yield. Evolutionary biology research uses comparative genomics to reveal the adaptive evolution of species and the evolution laws of gene families. For example, by comparing the genomes of polar fish and temperate fish, the molecular mechanisms related to cold resistance can be analyzed. In the field of clinical precision medicine, tumor heterogeneity analysis and rare disease diagnosis are core application scenarios. By comparing single-cell sequencing data with cross-species tumor models, the conserved mechanisms of tumor evolution can be revealed and assist in clinical decision-making. These application scenarios not only verify the value of the bioinformatics database, but also provide a direction for its continuous optimization, promoting the transformation of data resources towards scientific discovery and technological innovation.

4. Data Governance Strategies for Cross-Species Comparative Genomics

Cross-species comparative genomics reveals the laws of gene evolution, functional conservation and species adaptability mechanisms by integrating genomic data of multiple species. However, data heterogeneity, privacy protection requirements and analysis complexity pose challenges to data governance.

4.1 Data Lifecycle Management: Standardized Processes from Collection to Archiving

Data governance in cross-species comparative genomics needs to cover the entire process of data collection, storage, processing and archiving to ensure data quality, security and traceability.

In the data collection stage, a multi-source data integration mechanism needs to be established. Genomic sequencing data may come from public databases (such as NCBI SRA, ENA), cooperative laboratories or self-owned sequencing platforms. Information such as sample sources, sequencing strategies and experimental conditions needs to be recorded through a unified data submission interface and standardized metadata templates. For example, human genomic data need to be labeled with race, gender and disease status, and model organism data need to record strain, breeding environment and phenotypic data to support subsequent cross-species association analysis.

The data storage stage needs to adopt a hierarchical architecture. The original sequencing data (in FASTQ format) should be stored in high-performance distributed file systems (such as Lustre or Ceph) to support the rapid reading and writing of EB-level data; The comparison results (BAM/CRAM) and variant data (VCF) can be stored in object storage (such as MinIO), combined with compression algorithms to reduce storage costs. Meanwhile, a data version control system needs to be established to record the time, content and operator of each data update, so as to avoid the inability to reproduce the analysis results due to data overwriting.

The data processing stage needs to enhance efficiency through an automated assembly line. Build repeatable analysis processes using Nextflow or Snakemake, encapsulate steps such as sequencing quality assessment, genome assembly, annotation, and comparative analysis into modular tasks, and ensure environmental consistency through containerization technologies (such as Docker). For example, in cross-species gene family identification, OrthoFinder can be used to automatically perform homologous gene clustering, phylogenetic tree construction and functional enrichment analysis, reducing errors caused by manual intervention.

In the data archiving stage, attention should be paid to long-term preservation and accessibility. For the data of completed projects, it can be migrated to low-cost tape libraries or cloud storage (such as AWS Glacier), and a data summary report (including data volume, analysis results and citation suggestions) can be generated simultaneously for subsequent researchers to quickly assess the value of the data. In addition, a data destruction mechanism needs to be established to securely erase data that has exceeded its retention period or involves privacy risks, thereby preventing data leakage.

4.2 Cross-platform Collaboration Mechanism: Technological and Institutional Innovations to Break Down Data Silos

Cross-species comparative genomics requires the integration of data resources from multiple institutions worldwide and the achievement of data sharing and collaboration through technical tools and institutional design.

At the technical level, a unified data access interface and computing platform need to be constructed. Through RESTful API or GraphQL interfaces, genomic data, annotation information and variant directories scattered in different databases are integrated into a virtual data lake, supporting researchers to query multi-species data through a single entry point. For example, researchers can directly obtain the sequence, expression profile and evolutionary conservation information of the gene in humans, mice and zebrafish by inputting the gene name or coordinate range. Meanwhile, cross-platform analysis tools need to be developed, such as a cloud analysis environment based on JupyterHub, allowing researchers to upload their own data and conduct joint analysis with public data to avoid privacy risks caused by data transmission.

Data sharing needs to balance openness and security. For non-sensitive data (such as the genomes of model organisms), open access models (such as CC0 or CC-BY protocols) can be adopted and made globally available through data portals (such as Ensembl, UCSC Genome Browser); For data involving human genetic information or endangered species, a Data Use Agreement (DUA) needs to be established, requiring users to apply for permissions, commit not to use it for commercial purposes, and submit analysis reports regularly. In addition, sensitive data can be desensitized through differential privacy technology. For example, when sharing human genomic variation data, noise can be added to genotype frequencies, which retains statistical power while protecting individual privacy.

Cross-institutional collaboration requires the establishment of a trust mechanism and rules for benefit distribution. By using blockchain technology to record data sharing behaviors, including data access time, user identity and analysis purpose, data traceability and immutability can be achieved. Meanwhile, a data contribution assessment system needs to be established. Academic reputation or financial rewards should be allocated based on data quality, update frequency and sharing scope to encourage institutions to actively participate in data co-construction. For instance, in the Human Pan-Genome Project, each participant is granted the corresponding authorship rights based on the quantity and quality of the submitted variation data, promoting the synergy of global scientific research forces.

The ethical review and legal framework need to be improved at the institutional level. For the ethical issues that may be involved in cross-species comparative genomics (such as the comparison of genomes between humans and non-human primates), a multidisciplinary ethics committee needs to be established to review the research purpose, data usage methods and potential social impacts. Meanwhile, it is necessary to promote international legislative coordination, clarify data sovereignty, cross-border transmission rules and tort liability, and provide legal guarantees for global data collaboration.

5. International Practice and Localization Paths

5.1 Analysis of International Typical Models: The Balanced Practice of Open Sharing and Security Control

Internationally, the data governance of cross-species comparative genomics has formed a diversified model, the core of which lies in achieving a balance between data openness and privacy protection through technical tools and institutional design. In Europe and America, database construction led by public funds is representative. For instance, GenBank of the National Center for Biotechnology Information (NCBI) in the United States and ENA of the European Bioinformatics Institute (EBI) have

promoted the rapid sharing of global genomic data through mandatory data submission policies and open access protocols (such as CC0). This type of model relies on strict legal frameworks, such as the US Human Genetic Resources Management Act and the EU GDPR, to conduct cross-border transmission reviews and anonymization processing on data involving human genetic information. At the same time, it uses blockchain technology to record data access logs to ensure traceability. In addition, international alliance models (such as the Global Biodiversity Information Network GBIF) integrate species resource data from multiple countries through data sharing agreements among member states, supporting cross-regional ecological and evolutionary research.

Asian countries present differentiated paths. Japan has established a hierarchical data sharing system through the National Center for Genomic Medicine (NGM), classifying data into "public level", "restricted level" and "confidential level", and setting differentiated access rights for different levels. For example, clinical-grade genomic data is only available for internal use within cooperating institutions. Singapore, relying on its National Supercomputing Center, has built a cross-institutional collaboration platform based on federated learning, allowing participants to jointly model without sharing the original data and resolving the contradiction between data sovereignty and privacy protection. These models indicate that international practice is not a single path, but rather requires customized design in combination with regional scientific research capabilities, ethical norms and legal systems.

5.2 Localization Implementation Path: Technological and Institutional Innovation in Line with National Conditions

In the governance of cross-species comparative genomics data in China, it is necessary to base on scientific research demands and policy environments and explore implementation paths with local characteristics. From a technical perspective, the "federalization + edge computing" architecture can be prioritized for development. By deploying edge nodes in medical institutions and research institutes, local data processing and model parameter sharing can be achieved, reducing the risk of cross-border flow of sensitive data. For example, in the genomic research of rare diseases, a federated learning network covering the top three hospitals across the country can be constructed to jointly analyze multi-center sequencing data, and at the same time, patient privacy can be protected through homomorphic encryption technology. In addition, it is necessary to develop a genomic database that supports Chinese semantic retrieval, and combine NLP technology to analyze the gene names, phenotypic descriptions and experimental methods in Chinese scientific research literature, so as to improve the data utilization efficiency of local researchers.

At the institutional level, it is necessary to improve the data classification and grading standards as well as the sharing incentive mechanism. Referring to the "Data Security Law" and the "Personal Information Protection Law", a hierarchical directory of genomic data can be formulated, clearly defining the sharing scope and approval process of data at different levels. For instance, the genomic data of agricultural species can be classified as "open level" to support commercial breeding enterprises in using it in compliance. Classify human genetic resource data as "restricted level", and require users to submit ethical review certificates and commitments to share the results. Meanwhile, a quantitative assessment system for data contribution needs to be established. The volume of data submission, the volume of analysis tool development, and the volume of joint paper output should be included in the assessment indicators of research institutions to encourage their active participation in data co-construction. For example, in regional agricultural genome projects, special funds can be provided to units that share high-quality rice disease resistance gene data to promote the continuous accumulation of data resources.

The localization path also needs to pay attention to the construction of discourse power in international cooperation. China could lead the Asia-Pacific Regional Genomic Data Sharing Alliance to formulate data usage rules that are in line with the interests of developing countries. For instance, it could require multinational pharmaceutical companies to supply drugs to Chinese patients at preferential prices after developing them using genomic data from the Chinese population, thus avoiding the predicament of "data outflow - high-priced drugs flowing back". In addition, it is necessary to strengthen the brand building of domestic

databases. By increasing the frequency of data updates, the accuracy of annotations and the integration of analysis tools, international users can be attracted to use them actively, and the transformation from a "data provider" to a "data service provider" can be achieved.

6. Conclusion

Data governance in cross-species comparative genomics is a key support for modern life science research to move from single-point breakthroughs to systematic integration. Its core lies in achieving a dynamic balance between data openness and sharing and security and controllability through technological tool innovation and institutional design optimization. International practice shows that the open and shared model driven by public databases and legal frameworks in European and American countries, as well as the localized path combining hierarchical classification and federated learning in Asian countries, all reveal that data governance needs to be deeply compatible with regional scientific research ecosystems, ethical norms and legal systems. When promoting the governance of cross-species comparative genomics data, China needs to base itself on local demands and construct a three-in-one implementation framework of "technology - system - international cooperation" : At the technical level, the threshold for data sharing is lowered through federated learning, edge computing and Chinese semantic retrieval tools. At the institutional level, the efficiency of resource utilization is enhanced through classified and graded standards, data contribution assessment and sharing incentive mechanisms. In international cooperation, it is necessary to strive for the right to speak in global data governance by leading regional alliances, formulating fair rules and strengthening the service capabilities of local databases. In the future, with the continuous decline in the cost of genomic sequencing and the explosive growth of multi-omics data, data governance will shift from "tool optimization" to "ecological reconstruction". It is necessary to further explore the automated governance framework driven by privacy computing, blockchain evidence storage and artificial intelligence, and promote global scientific research collaboration under the premise of safeguarding data sovereignty. To provide sustainable data infrastructure support for analyzing the laws of life evolution, conquering human diseases and ensuring food security.

References

[1] Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., ... & Paten, B. (2023). A draft human pangenome reference. Nature, 617(7960), 312-324.

[2] Gupta, P. K. (2022). Earth Biogenome Project: present status and future plans. Trends in Genetics, 38(8), 811-820.

[3] Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., ... & Zhang, G. (2022). The Earth Biogenome Project 2020: Starting the clock. Proceedings of the National Academy of Sciences, 119(4), e2115635118.

[4] Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. Nature genetics, 52(7), 646-654.

[5] Martinez-Morales, J. R. (2016). Toward understanding the evolution of vertebrate gene regulatory networks: comparative genomics and epigenomic approaches. Briefings in Functional Genomics, 15(4), 315-321.

[6] Stephens, Z. D., Lee, S. Y., & Faghri, F. (2015). Big Data: Astronomical or Genomical?. PLoS Biology. PLoS Biology, 13(7), e1002195.

[7] Modi, A., Vai, S., Caramelli, D., & Lari, M. (2021). The Illumina sequencing protocol and the NovaSeq 6000 system. In Bacterial pangenomics: methods and protocols (pp. 15-42). New York, NY: Springer US.

[8] Bryce, A. H., Egan, J. B., Borad, M. J., Stewart, A. K., Nowakowski, G. S., Chanan-Khan, A., ... & McWilliams, R. R. (2017). Experience with precision genomics and tumor board, indicates frequent target identification, but barriers to delivery. Oncotarget, 8(16), 27145.

[9] European Bioinformatics Institute: Birney Ewan 3 Goldman Nick 3 Kasprzyk Arkadiusz 3 Mongin Emmanuel 3 Rust Alistair G. 3 Slater Guy 3 Stabenau Arne 3 Ureta-Vidal Abel 3 Whelan Simon 3, Research Group in Biomedical Informatics Abril Josep F. 5 Guigó Roderic 5 Parra Genís 5, Bioinformatics Agarwal Pankaj 6, National Center for Biotechnology Information Agarwala Richa 7 Church Deanna M. 7 Hlavina Wratko 7 Maglott

Donna R. 7 Sapojnikov Victor 7, Department of Mathematics Alexandersson Marina 8 Pachter Lior 8, Division of Medical Genetics Antonarakis Stylianos E. 9 Dermitzakis Emmanouil T. 9 Reymond Alexandre 9 Ucla Catherine 9, ... & Department of Biology Lander Eric S. lander@ genome. wi. mit. edu 2 46 b. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915), 520-562.

[10] Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome research, 27(5), 665-676.